# Methods and Techniques for Generation and Integration of Web Ontology Data

A Thesis Submitted for the Degree of
Doctor of Philosophy

By

*Chao Wang*

in

FACULTY OF INFORMATION TECHNOLOGY
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
JUNE 2007

UNIVERSITY OF TECHNOLOGY, SYDNEY
FACULTY OF INFORMATION TECHNOLOGY

The undersigned hereby certify that they have read this thesis entitled "**Methods and Techniques for Generation and Integration of Web Ontology Data**" by **Chao Wang** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of **Doctor of Philosophy**.

Dated: June 2007

Research Supervisors: _____
Dr Jie Lu

_____
Dr Guangquan Zhang

# CERTIFICATE

Date: **June 2007**

Author: **Chao Wang**

Title: **Methods and Techniques for Generation and Integration of Web Ontology Data**

Degree: **Ph.D.**

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.

Signature of Author

# Acknowledgements

I would like to express my sincere and deep gratitude to my principal supervisor, Jie Lu, for her continuous encouragement, advice, help and invaluable suggestions. Without her support, I couldn't even start the PhD study. She is such a nice, generous, helpful and kindhearted person. I feel really happy, comfortable and unconstrained with her during my PhD study. I owe my research achievements to her professional supervision.

Many thanks are also due to my co-supervisor, Guangquan Zhang, for his valued suggestions and constant support, and for the numerous conversations with him.

I wish to thank my fellow research students and the staff of the Faculty of Information Technology, University of Technology, Sydney (UTS). Their various assistance and advice are of great benefit to this study.

I appreciate the financial support from both the UTS International Research Scholarship (IRS) and the Faculty of Information Technology Scholarship. I appreciate the travel support for attending the international conferences which I received from the Faculty of Information Technology and the UTS Vice-Chancellor's Conference Fund.

I would like to express my heartfelt appreciation to my parents for their constant support, encouragement and care over all these years. Also, I would like to thank my wife Kei Shum for her support and love during my research in the past years.

*To my parents, Limin Wang and Weijuan Feng*

# Table of Contents

# List of Tables

# List of Figures

# Abstract

Data integration over the web or across organizations encounters several unfavorable features: heterogeneity, decentralization, incompleteness, and uncertainty, which prevent information from being fully utilized for advanced applications such as decision support services. The basic idea of ontology related approaches for data integration is to use one or more ontology schemas to interpret data from different sources. Several issues will come up when actually implementing the idea: (1) How to develop the domain ontology schema(s) used for the integration; (2) How to generate ontology data for domain ontology schema if the data are not in the right format and to create and manage ontology data in an appropriate way; (3) How to improve the quality of integrated ontology data by reducing duplications and increasing completeness and certainty.

This thesis focuses on the above issues and develops a set of methods to tackle them.

First, a key information mining method is developed to facilitate the development of interested domain ontology schemas. It effectively extracts from the web sites useful terms and identifies taxonomy information which is essential to ontology schema construction. A prototype system is developed to use this method to help create domain ontology schemas.

Second, this study develops two complemented methods which are light weighted and more semantic web oriented to address the issue of ontology data generation. One method allows users to convert existing structured data (mostly XML data) to ontology data; another enables users to create new ontology data directly with ease.

In addition, a web-based system is developed to allow users to manage the ontology data collaboratively and with customizable security constraints.

Third, this study also proposes two methods to perform ontology data matching for the improvement of ontology data quality when an integration happens. One method uses the clustering approach. It makes use of the relational nature of the ontology data and captures different situations of matching, therefore resulting in an improvement of performance compared with the traditional canopy clustering method. The other method goes further by using a learning mechanism to make the matching more adaptive. New features are developed for training matching classifier by exploring particular characteristics of ontology data. This method also achieves better performance than those with only ordinary features. These matching methods can be used to improve data quality in a peer-to-peer framework which is proposed to integrate available ontology data from different peers.